한국학교보건교육학회지 제 11권 1호 (2010. 6) pp. 67～78
The Journal of Korean Society for School Health Education, Vol. 11 Iss. 1 (June, 2010)

청소년 건강관련 공개자료 접근 및 활용에 관한 고찰

이재은 · 성정혜† · 이원재* · 문인옥**

Department of Epidemiology & Biostatistics in Jackson State University, USA,
*경원대학교 의료경영학과 · **이화여자대학교 보건관리학과

Access to and Utilization of the Open Source Data-related to Adolescent Health

Jae Eun Lee · Jung Hye Sung† · Won Jae Lee* · In Ok Moon**
Department of Epidemiology & Biostatistics in Jackson State University, USA,
*Department of Healthcare Management in Kyungwon University,
**Department of Health Education & Management in Ewha Womans University

ABSTRACT

Background & Objectives: Current trend is that funding agencies require investigators to share their data with others. However, there is limited guidance how to access and utilize the shared data. We sought to determine what common data sharing practices in U.S.A. are, what data-related to adolescent health are freely available, and how we deal with the large dataset adopting the complex study design.

Methods: The study included only research data-related to adolescent health which was collected in USA and unlimitedly accessible through the internet. Only the raw data, not aggregated, was considered for the study. Major keywords for web search were "adolescent", "children", "health", and "school".

Results: Current approaches for public health data sharing lacked of common standards and varied largely due to the data's complex nature, large size, local expertise and internal procedures. Some common data sharing practices are unlimited access, formal screened access, restricted access, and informal exclusive access. The Inter-University Consortium for Political and Social Research and the Center for Disease Control and Prevention were the best data depository. "Data on the net" was search engine for the website providing data freely available. Six datasets related to adolescent

접 수 일 : 2010년 1월 6일, 수 정 일 : 2010년 1월 26일, 채 택 일 : 2010년 2월 8일
† 교신저자 : Jung Hye Sung(350 W. Woodrow Wilson Department of Epidemiology and Biostatistics School of Health Sciences College of Public Service Jackson State University, USA.)
Tel: 601-979-8792, Fax: 601-979-8802, E-mail: Jung.h.lee@jsums.edu

health freely available were identified. The importance and methods of incorporating complex research design into analysis was discussed.

Conclusion: There have been various attempts to standardize process for open access and open data using the information technology concept. However, it may not be easy for researchers to adapt themselves to this high technology. Therefore, guidance provided by this study may help researchers enhance the accessibility to and the utilization of the open source data.

Key words: Data sharing, Open source data, Open source statistical software, Secondary data

Ⅰ. 서론

정부기관에서 발주하는 조사연구는 국민의 세금으로 형성된 기금을 통해 진행된다. 따라서 수집된 조사연구자료는 조사를 주도한 특정 그룹이나 특정인의 소유가 아닌 관심을 가지고 있는 일반 대중이나 연구자들과 공유해야 할 공공재(Public goods)로서의 성격을 지니고 있다고 할 수 있다. 자료공유란 원래 계획된 것과는 다른 목적으로 자료를 이용하고자 하는 사람들을 위해 자료를 공시(公示)하는 것을 의미한다. 자료공유에는 연구발주자(정부 혹은 연구기금 제공자), 자료생산자(원 연구자) 그리고 자료 사용자의 이해관계가 얽혀있다. 즉, 연구발주자는 자료 공유를 통해 학문의 발전에 기여하고 국민의 알 권리를 충족시키며, 중복연구를 방지하여 기금을 효율적으로 관리할 수 있기 때문에 자료를 공유하길 원하고 있으며, 자료 사용자는 자료수집을 위한 시간, 비용, 노력을 절약할 수 있어 납세자로서의 자료에 대한 권리를 주장하고 있다. 자료생산자인 원 연구자는 자료의 소유권을 잃게 되어 저작권이나 연구결과를 독점할 수 있는 기회를 상실하게 될지도 모른다는 우려를 하고 있어 공개를 꺼리고 있다. 따라서, 자료공유에 있어서 가장 커다란 장애 요인은 자료생산자라고 할 수 있다.

자료공유는 학문의 발전을 진작시킨다는 면에서 사회적으로 가치 있는 일이다. 이러한 혜택은 기금제공자나 자료사용자에게만 국한된 것이 아니다. 자료생산자 역시 많은 혜택을 누릴 수 있다. 즉, 외부연구자들과의 협동연구를 통해 논문발표의 기회를 늘릴 수 있고, 자료생산자에 의해 발표된 주요 연구에 대한 인용횟수를 늘릴 수 있다. 인용지수는 연구자의 능력 및 발표된 논문의 질을 평가하는 주요한 기준이며 연구자나 연구기관의 명성에 영향을 주므로 인용자수가 높아진다면 지속적인 연구사업기금을 획득하는데 유리한 위치를 차지할 수 있다(Piwowar 등, 2008). 최근의 연구에 의하면 데이터를 공유하는데 있어서 장애요인은 자료의 분석 및 저작과 직접적으로 관련 있는 자료 소유권이 아니라, 공유를 위한 기술적인 측면(IT기술, 자료의 표준화 등), 관리적인 측면(접근수준, 인권보호 등) 그리고 금전적인 측면(자료관리 비용, 관련 문서 작성을 위한 비용 등) 등이 더 큰 문제인 것으로 나타났다(Chandramohan 등, 2008). 이는 자료공유가 더 이상 거스를 수 없는 학계 전반의 대세이며, 단지 공유를 위해 지불해야 하는 추가적인 비용을 어떻게 처리할 것인가가 남아있는 숙제라는 것을 의미한다.

자료공유는 새로운 이슈가 아니다. 이미 다른 연구분야에서는 많은 자료들이 무료로 공유되고 있다. 특히 유전학, 환경과학, 천체물리학 등에서는 무료로 접근 가능한 자료들이 무수히

존재한다. 하지만 보건분야에서 데이터공유 문화가 정착된 것은 그리 오래 되지 않는다. 이는 보건관련 연구 조사 데이터에는 조사대상자에 대한 예민한 정보를 포함하고 있어 인권이 침해될 소지가 높기때문인 것으로 생각된다. 미국의 경우, 2003년 10월 1일부터 미국국립보건원(National Institute of Health 이하 NIH)에서 발주하는 조사연구사업 중 직접비용이 연간 50만 달러 이상인 경우에는 데이터 공유에 대한 계획을 연구계획서에 포함하거나, 데이터 공유가 불가능할 경우에는 그 이유를 명기할 것을 의무화하고 있다. NIH는 데이터 공개 시기를 주요 논문이 학술지 게재가 확정된 시기로 규정하고 있다. Robert Wood Johnson Foundation과 같은 사설 재단에서도 데이터공유와 관련한 의무규정을 명시하고 있다. 또한, 일반 학술지 차원에서도 발표될 논문에 사용된 원자료(raw data)를 공유할 것을 권고하고 있는 실정이다(Savage and Vickers, 2009). 이러한 학계 전반에 걸친 데이터 공유문화를 조성하기 위한 노력에 의해 현재 많은 자료들이 다양한 경로를 통해 공유되고 있다.

2차 자료는 자료이용자의 연구취지에 부합하는 모든 변수를 포함하고 있지 않으며, 표본규모나 표본추출이 연구목적과 일치하지 않는 단점이 있을 수 있음에도 불구하고, 2차 자료에 대한 접근은 학위논문을 준비하는 대학원생이나 데이터 수집을 위한 시간이 부족한 젊은 학자들에게 있어서 여전히 매력적인 연구 대안일지 모른다. 그리고 대학강단에서 강의를 준비하는 교수들에게도 2차 자료는 좋은 교육자료로 이용될 수 있을 것이다. 하지만 국내 사정은 원자료의 공유가 제한되어 있는 실정이고, 공유된다고 하더라도 유료인 경우가 많다. 더구나 무료로 공유되는 데이터에 대한 접근법과 활용에 관한 상세한 지침서나 연구는 제한되어 있다. 따라서, 본 연구는 현재 미국에서 데이타 공유를 위해 어떠한 방법을 사용하고 있는지, 무료로 접근 가능한 청소년건강관련 자료에는 무엇이 있는지, 그리고 자료처리 시 고려해야 할 점은 무엇인지를 살펴봄으로써 2차 자료를 이용하고자 희망하는 모든 연구자들에게 자료에 대한 접근성을 높여 주고 데이터 활용을 촉진시키고자 하는데 그 목적이 있다.

Ⅱ. 연구방법

일반적으로 데이터에는 구체적인 연구 목적을 달성하기 위해 이미 수립된 계획에 따라 수집하는 자료(연구조사 자료)가 있는가 하면 특별한 연구목적 없이 수집하는 자료(예, 암등록 데이터, 센서스 테이터, 환자기록 등)가 있다. 본 연구에서는 연구조사자료(research data)만을 그 대상으로 한정하였다. 연구조사자료란 연구목적이 조사결과를 통해 입증될 수 있는지 여부를 판단하기 위해 과학계에서 통상적으로 받아들여진 방법을 통해 수집된 숫자나 문자로만 이루어진 모든 사실기록물(recorded factual material)을 의미한다. 따라서 조사자료를 바탕으로 가공된 결과자료(예를 들면, 예비분석결과, 발표준비를 위한 논문 초고), 조사 계획서, 물리적인 대상물(실험표본, 오디오/비디오 테잎), 법에 의해 보호받고 있는 정보나 개인의 인권을 침해할 소지가 있는 자료 등은 본 연구에서 제외하였다. 요약자료(aggregated data)는 연구에서 제외되었고 원자료(raw data)만을 포함하였다. 하지만 전수조사일 경우 최소단위(block group level, census track 등)에서 요약되는 것이 일반적이므로 제외시키지 않았다.

자료생산자의 권리를 보호하기 위해 데이터 접근에 여러 가지 제한을 두는 경우가 일반적인데 본 연구에서는 완전히 무료로 개방된 원자료만을 고려했다. 또한 자료를 공유하기 위해

여러 매체를 사용하고 있으나 본 연구에서는 인터넷을 통해 접근가능한 자료만을 그 대상으로 하였다. 공개자료의 범위는 미국 내의 자료로 한정하였으나 보충설명을 위해 한국과 WHO의 자료개방 사례를 제시하였다. 청소년기의 질병사망률은 낮을지라도 성인들의 조기사망률의 많은 부분이 청소년기의 비 건강행위(예, 흡연, 음주, 잘못된 식습관 등)에 기인한다는 사실은 이미 여러 연구를 통해 잘 알려져 있다. 따라서, 최근 청소년의 건강행태 등이 중요한 사회적 문제로 부각되고 있는 시점에 즈음하여, 본 연구는 청소년 건강증진을 통해 국민건강증진을 도모하고자 하는 연구자들에게 도움을 주고자 청소년 건강관련 자료만을 그 대상으로 하였다. 청소년 건강관련 자료를 수집하기 위해 “adolescent”, “children”, “health”, “school” 등의 키워드를 사용하였다.

Ⅲ. 결과 및 토론

1. 보건분야 자료공유를 위한 접근법

현재 보건분야 자료공유를 위한 접근방법에는 공통된 기준이 없다. 이는 데이터 자체가 지니고 있는 복잡한 특성(예, 개인의 사생활에 피해를 줄 수 있는 개연성이 높은 자료는 접근도를 결정하는 중요한 요인) 때문이기도 하고, 자료 공유에 대한 내부절차가 다양하기 때문이기도 하다. 하지만 데이터에 대한 접근 정도에 따라서 다음과 같은 방법들이 시행되고 있다.

- 무제한 접근(unlimited access): 자료이용자가 자격요건을 갖추지 않았더라도 무료로 자료를 무제한 내려 받을 수 있다. 간단한 등록을 요구하는 경우도 있지만 자격기준을 평가하려는 목적이 아니다. 대학기관에서 수집한 데이터나 혹은 정부기관이 주요 연구자일 경우 이와 같은 방법을 사용하여 자료를 공유하고 있다.
- 공식적 자격요건 접근(formal screened access): 외부자료 이용자가 자료를 취득하기 위해서는 지정한 양식에 기입하고 연구계획서를 제출해야 한다. 내부절차에 따라 제출된 연구계획서를 평가하여 통과된 경우에 연구를 위해 필요한 데이터(customized dataset)만을 이용자에게 제공하는 경우를 말한다. 대형 심장병 역학조사인 Framingham Heart Study, Jackson Heart Study 등이 이런 방법으로 자료를 공유하고 있다. 자료신청 절차 및 방법은 통상적으로 각 연구의 홈페이지에 자세히 설명되어 있다.
- 제한적 접근 (restricted access): 자료이용 전에 연구 계획서를 제출하여 통과되어야 하며, 자료이용자가 자료를 보관하고 있는 장소를 직접 방문하여 관계자의 엄격한 감독 하에 자료를 분석하고 그 결과만을 가지고 나올 수 있다. 예를 들면, National Health and Nutrition Examination Survey의 유전자 데이터를 분석하기 위해서는 워싱턴 DC에 있는 미건강통계센터를 직접 방문해야 하며 센터 내의 분석실에서만 분석할 수 있고 외부에는 분석결과만을 가지고 나올수 있다.

이 밖에도, 비공식 배타적 접근(informal exclusive access)방법으로 간단한 구두계약을 통해 개인적인 친분이 있는 연구자와 자료를 공유하는 경우도 있다. 연구금액이 작은 소형 연구사업의 경우에는 자료공유에 대한 계획을 요구하고 있지 않기 때문에 이와 같은 배타적인 방법을 사용하고 있다. 또한, 사설기금에 의해 진행된 연구자료의 경우 이와 같은 방법을 사용하는 경우가 있었다. 예를 들면, Robert Wood Johnson기금을 통해 1994년에서 2000년

대 초까지 진행된 Harvard College Alcohol Study는 자료수집 후 몇 년 간은 비공식적이고 배타적인 방법으로 가까운 연구자들간에만 자료를 공유하였다. 현재는 무제한 접근을 통해 자료가 공유되고 있다. 또한 복수의 접근방법을 통해 자료가 공유되는 경우도 있다. 예를 들면, The National Longitudinal Study of Adolescent Health (Add Health: http://www.cpc.unc.edu/projects/addhealth)는 무제한 및 제한적 접근을 동시에 사용하고 있다.

자료공유를 위해 인터넷 매체를 사용하는 것이 일반적인데, 인터넷을 통한 데이터 공유에는 다음과 같은 두 가지의 방법이 있다.

- 데이터 보관소(data depository): 회원들이 자발적으로 원자료와 관련 서류(설문지, 코딩북, 자료수집 절차 등)를 자료은행에 예치하고 회원 상호간에 데이터를 공유하는 방법이다. 검색기능(search engine)을 통하여 관련 자료를 검색할 수 있으며 방대한 종류의 원자료를 구할 수 있다. ICPSR(The Inter-University Consortium for Political and Social Research)은 데이타 보관소로서 가장 대표적이며 뛰어난 것으로 알려졌다. ICPSR은 1962년에 설립된 미국 최대 사회과학데이터보관소로서 세계 각 국에서 수집된 7,500여 개의 조사자료(survey data)를 소장하고 있다. 현재 전 세계 700여 개의 대학이 컨소시엄으로 구성되어 있다. 비회원이라 할지라도 대부분의 보건/범죄 관련 데이터는 무료로 제공받을 수 있다. ICPSR은 <Table 1>과 같이 별도의 보건관련 데이터보관소를 운영하고 있다. 이 밖에도 소규모 연구 네트워크에서도 회원상호간의 데이터 공유를 위해 데이터 보관소를 운영하는 것이 대세인 듯하다. 예를 들면, 미국 대도시에 위치한 한인들을 위한 건강 서비스 및 연구 센터들은 수집된 자료를 비교분석하기 위해 데이터 보관소 운영을 계획하고 있다. 한편 국내의 대표적인 자료보관소는 한국 사회과학 데이터센터(Korean Social Science Data Center: KSDC)를 들 수 있다. KSDC는 ICPSR의 한국측 국가회원(National Membership)으로 국내 대학 및 연구기관에 서비스하고 있다. KSDC가 보유하고 있는 국내 보건관련 자료로는 19개의 여론조사 자료(예, 보건의료 서비스에 대한 소비자 만족도 조사, 건강상담실 운영에 대한 설문조사, 가정노인보호서비스 제도 도입 방안에 대한 설문조사 등)와 2개의 통계자료(예, 국민영양, 보건의료인력 및 시설)가 있었다. 하지만 여론조사 자료는 1990년대에 수집된 자료가 대부분이어서 활용성 측면에서 문제가 있을 것으로 사료된다.
- 개별 연구 홈페이지(study website): 각 개별 연구기관의 홈페이지를 통해 데이터를 공유하는 경우도 흔한 일이다. 가장 대표적인 곳인 미국 질병관리본부(Center for Disease Control and Prevention)산하 건강통계센터의 홈페이지에는 8개의 대형 조사자료를 무료로 제공하고 있다. 이 밖에도 BRFSS(Behavioral Risk Factor Surveillance System)와 YRBSS(Youth Risk Behavioral Surveillance System)등 대형 데이터가 질병관리본부 홈페이지를 통해 무료로 공유되고 있다. 세계보건기구 홈페이지에서도 훌륭한 건강관련 자료를 제공하고 있다. “Data on the Net”은 자료를 보유하고 있는 홈페이지의 주소를 검색하기 위한 사이트로 현재 156개의 자료제공 홈페이지 주소가 등록되어 있다.

<Table 1> Major Health-related Data Depository

Depository name	Description	Web site
Child Care and Early Education Research Connections	Research Connections (RC) offers a comprehensive, easily searchable collection of more than 10,000 resources from the many disciplines related to child care and early education. The site offers the most current publications, as well as links to child care policy statements.	http://www.researchconnections.org/
Data Sharing for Demographic Research (DSDR)	DSDR provides resources to demographic data producers and users, including confidentiality and disclosure review, restricted data contract development and data dissemination, a searchable index of important demography and population study data, and a catalogue of publications using data indexed.	http://www.icpsr.umich.edu/DSDR/index.html
Health and Medical Care Archive (HMCA)	HMCA preserves and disseminates data collected by research projects funded by The Robert Wood Johnson Foundation, the nation's largest philanthropy devoted exclusively to improving the health and health care of all Americans.	http://www.icpsr.umich.edu/HMCA/index.html
National Archive of Computerized Data on Aging (NACDA)	NACDA acquires and provides access to data relevant to gerontological research with an emphasis on major issues of scientific and policy relevance.	http://www.icpsr.umich.edu/NACDA/index.html
Minority Data Resource Center (MDRC)	MDRC is a membership-funded archive whose purpose is to provide data resources for the comparative analysis of issues affecting racial and ethnic minority populations in the United States. We offer streamlined access to existing ICPSR data and to newly acquired studies that are relevant to the study of immigration, place of origin, ancestry, ethnicity, and race in the United States. Access to MDRC data is available to anyone at an ICPSR member university or institution.	http://www.icpsr.umich.edu/MDRC/
Substance Abuse and Mental Health Data Archive (SAMHDA)	SAMHDA provides public data access and online analysis for important substance abuse and mental health data collections. The project offers variable-level searching, an archive of survey instruments, related literature for data collections, a listserv, disclosure analysis, and traditional data products. SAMHDA was established at ICPSR in 1995 by the Office of Applied Studies (OAS), Substance Abuse and Mental Health Services Administration (SAMHSA).	http://www.icpsr.umich.edu/SAMHDA/index.html

<Table 2> Major Adolescent Health-Related Data that are Being Shared Freely Through Internet

Study Name	Principal investigator	Data collection period	Objectives and Methods	Major subjects	Website
Monitoring the Future	Survey Research Center in the Institute for Social Research at the University of Michigan.	From 1975 to current	This is the annual survey in this series that explores changes in important values, behaviors, and lifestyle orientations of contemporary American youth. Students are randomly assigned to complete one of six questionnaires, each with a different subset of topical questions, but all containing a set of "core" questions on demographics and drug use. There are about 1,400 variables across the questionnaires.	alcohol, attitudes, drug use, family life, high school students, life plans, lifestyles, religious attitudes, social behavior, social change, tobacco use, values, youths	http://monitoringthefuture.org/
2007 Health Tracking Household Survey	Center for Studying Health System Change	2007; It is the successor to the Community Tracking Study (CTS).	Households were selected using random-digit dialing techniques. The survey obtained information about every adult and one child under age 18 (if any) in each family insurance units (FIU). In FIU's with more than one child, one child was randomly selected for inclusion in the survey.	doctor visits, health care access, costs, facilities, utilization, health insurance, mental health services, physician choice, physician-patient relationship,	http://www.hschange.org/
National Health Interview Survey	United States Department of Health and Human Services. National Center for Health Statistics	Annual survey since 1957	The purpose of the National Health Interview Survey (NHIS) is to obtain information about the amount and distribution of illness, its effects in terms of disability and chronic impairments, and the kinds of health services people receive. It is a cross-sectional household interview survey. The sampling plan follows a multistage area probability design that permits the representative sampling of households and noninstitutional group quarters (e.g., college dormitories).	assistive devices, child health, disabilities, doctor visits, health behavior, access, services, utilization, health status, hospitalization, illness, immunization, injuries, mental health, physical disabilities, poisoning	http://www.cdc.gov/nchs/nhis.htm

(Continue)

Study Name	Principal investigator	Data collection period	Objectives and Methods	Major subjects	Website
National Survey of Children's Health	United States Department of Health and Human Services. Centers for Disease Control and Prevention. National Center for Health Statistics	2003, 2007	The National Survey of Children's Health (NSCH) was designed to produce national- and state-specific prevalence estimates for a variety of physical, emotional, and behavioral health indicators and measures of children's experiences with the health care system. There were 102,353 children, selected through a random-digit-dial sample of households with children under 18 years of age, who were selected from each of the 50 States and the District of Columbia. One child was randomly selected from all children in each identified household to be the subject of the survey. The respondent was the parent or guardian who knew the most about the child's health and health care.	child health, health care access, needs assessment, children, adolescents, mental health, emotional states, stress, child development, everyday life, social behavior, family structure, health insurance, neighborhood conditions, welfare services	http://www.nschdata.org/Content/Default.aspx
Youth Risk Behavior Surveillance	the Centers for Disease Control and Prevention (CDC) and state, territorial, tribal, and local governments		The Youth Risk Behavior Surveillance System (YRBSS) monitors priority health-risk behaviors and the prevalence of obesity and asthma among youth and young adults. The YRBSS includes a national school-based survey conducted by the Centers for Disease Control and Prevention (CDC) and state, territorial, tribal, and local surveys conducted by state, territorial, and local education and health agencies and tribal governments.	unintentional injuries and violence; tobacco use; alcohol and other drug use; sexual behaviors that contribute to unintended pregnancy and STDs infection; unhealthy dietary behaviors; physical inactivity; obesity; asthma.	http://www.cdc.gov/HealthyYouth/yrbs/index.htm
National Survey on Drug Use and Health (formerly National Household Survey on Drug Abuse)	United States Department of Health and Human Services. Substance Abuse and Mental Health Services Administration. Office of Applied Studies	Annual since 1971	The National Survey on Drug Use and Health (NSDUH) series (formerly titled National Household Survey on Drug Abuse) primarily measures the prevalence and correlates of drug use in the United States. A multistage area probability sample for each of the 50 states and the District of Columbia has been used since 1999.	addiction, alcohol, alcohol abuse, alcohol consumption, drinking behavior, mental health, mental health services, prescription drugs, smoking, substance abuse, substance abuse treatment	http://www.oas.samhsa.gov/nhsda.htm

2. 청소년 건강 관련 공개자료 현황 (open source data)

비록 여러 개의 데이터 보관소가 운영되고 있지만 많은 자료들이 복수의 보관소에서 중복적으로 검색되고 있다. 특히 ICPSR에는 다른 보관소에서 검색되지 않은 데이터도 있어 ICPSR이 모 보관소(mother depository)로서의 역할을 하고 있었다. 하지만 개별 홈페이지를 통해 공유되고 있는 데이터는 데이터 보관소에서 검색되지 않는 경우가 많았다. 따라서 완벽한 자료검색을 위해서는 데이터 보관소 뿐만 아니라 개별연구 홈페이지에 대한 검색도 병행해야 한다. 특히 ICPSR과 미국 질병관리본부 홈페이지에 대한 검색을 통한다면 훌륭한 청소년 건강 관련 데이터를 용이하게 취득할 수 있을 것이다. 데이터 보관소나 개별 연구 홈페이지를 통해서 무료로 공유되는 청소년 건강관련 원자료와 관련 문서들은 무수히 많았다. 이를 모두 적시한다는 건 불가능하며 무의미하다고 생각되어, <Table 2>와 같이 인터넷을 통해 무료로 데이터가 공유되고 있는 대표적인 청소년 건강 연구들만을 요약 정리하였다.

3. 데이터를 내려 받을 때 유의할 점

먼저 데이터를 내려 받기 전에 이용약관(term of use)을 꼼꼼하게 읽어보고 제시한 규정을 준수할 것에 동의해야 한다. 무료로 접근 가능한 데이터는 까다로운 규정을 제시하지 않는 것이 보통이다. 예를 들면, 논문제목에 조사연구명을 포함하라든가, 발표된 논문은 데이터 제공 홈페이지에 등록하라는 등 간단한 의무조항을 포함하는 것이 일반적이다. 의무불이행 시 처벌은 약하지만 건전한 데이터 공유문화를 위해 의무조항을 준수하는 것이 바람직하리라 사료된다. 이용약관에 동의한 다음, 사용가능한 통계패키지 프로그램이 무엇인지를 파악하고 적합한 포맷의 데이터를 내려 받아야 한다. 공유데이터는 SAS 혹은 SPSS프로그램에서 읽어낼 수 있게 포맷되어 있는 경우가 보통이다. 하지만 위의 두 패키지 프로그램을 가지고 있지 않을 경우에는 ASCII파일로 내려받은 후, 사용하고 있는 통계프로그램으로 읽어야 한다. ASCII파일은 어떠한 통계프로그램에서도 읽을 수 있다. 데이터를 내려 받을 때 코드북, 분석지침서, 데이터 수집방법, 설문지 등의 관련 문서도 함께 내려 받아야 한다.

4. 통계처리 시 유의할 점

통계처리 전에 분석지침서를 충분히 숙지해야 한다. 분석지침서는 정기적으로 갱신되는 경우도 있기 때문에 해당 데이터 홈페이지에 접속하여 최신 지침서를 내려 받아야 한다. 대부분의 대형 연구사업은 복잡한 표본추출방법을 사용하고 있는데 이를 분석에 고려하지 않을 경우 부정확한 추정치를 얻거나 type-1오류를 범하게 된다. 복잡한 표본추출방법을 사용하고 있음에도 이를 분석에 반영하지 않으면 그 분석결과를 인정하지 않는 것이 미국학계의 경향이며, 학술논문 심사위원으로부터 재분석을 요구 받거나 이에 응하지 않을 경우 학술지에 채택되지 않는 경우가 대부분이다. 표본추출을 채택한 연구 조사는 응답자가 선택될 확률 및 비응답률이 각 기초표본단위(primary sampling unit; 예, 가정)에 따라 다르며, 표본과 실제의 주요 인구 구성요소가 차이가 나는 것이 일반적이다. 이를 보정하기 위해서는 데이터에 가중치를 부여해야 한다. 따라서 불편추정치(unbiased estimate)를 얻기 위해서는 표본가중치(sampling weights)를 분석모형에 포함시켜야 한다. 정확한 추정치 자체가 연구목적인 경우(예, 전국의 중학교 학생들의 흡연율이 얼마인지)에는 반드

시 가중데이터(weighted data)를 사용해야 한다. 여러 해에 걸쳐 수집된 데이터(multi-year data)을 통합하여 분석할 때에는 새로운 가중치를 써야 하는 경우도 있다. 예를 들면, NHANES는 정기적으로 분석지침을 갱신하고 있으며 새로운 가중치에 대한 정보도 발표한다.

가중치와 함께 표본군집(sample clusters)도 동시에 분석에 고려해야 한다. 표본가중치는 추정 값에 영향을 미치지만 표본군집은 standard error(SE)와 유의도에 영향을 미친다. 즉, 표본군집을 분석에 고려하였을 경우 SE가 늘어나 유의도가 감소하며 type-1 오류의 가능성을 배제할 수 있어 보다 확고하고 보수적인 추정(robust and conservative inference)을 가능하게 한다. 전통적인 통계기법으로는 표본군집효과를 고려할 수 없기 때문에 특별한 방법(예, Taylor series linearization method, balanced repeated replications 등)을 적용해야 하는데, 이는 SAS, Sudaan, Stata 등의 특정한 통계 프로그램을 통해서 가능하다. 위의 세 가지 통계 프로그램은 가격이 비싸기 때문에 학생들이나 초보연구원들에게는 적합하지 않을지도 모른다. 하지만 R Project, AM, WesVar, Epi-info 등의 무료 통계프로그램을 사용한다면 표본군집을 분석에 효과적으로 고려할 수 있다. R Project와 Epi-info는 SAS와 마찬가지로 다목적용 통계프로그램으로 응용력은 높으나 이들을 활용하기 위해서는 높은 수준의 프로그래밍 능력을 가지고 있어야 한다. 반면 AM과 WesVar는 간단히 몇 개의 메뉴 바를 클릭하면 결과를 얻어낼 수 있으며, 여러 종류의 포맷 데이터를 읽어들일 수 있는 기능을 갖추고 있어 초보자들에게 적합한 통계 프로그램이라고 할 수 있다.

Ⅳ. 결론

2차 데이터 사용자는 자료수집에 요구되는 시간과 경비를 절약하여 이론개발을 위한 가설검증연구(hypothesis test studies)에 전념할 수 있다는 장점이 있다. 그러나 자료공유가 대세이긴 하지만 부정적인 측면도 무시할 수는 없다. 즉, 과학자들에게 많은 시간이 요구되는 연구계획서를 작성하고 데이터를 수집하는 등의 일에 동기부여를 할 수 없기 때문에 궁극적으로 1차 자료 생산이 감소할 수도 있다. 하지만 젊은 학자나 학위논문을 준비하는 학생들에게 2차 자료는 여전히 매력적인 대안이라고 할 수 있다.

본 연구는 현재 미국에서 자료공유가 어떻게 진행되고 있는지, 자료공유를 위한 방법에는 무엇이 있으며, 무료로 공유되는 청소년 건강관련 자료에는 어떤 것이 있고 어떻게 접근할 수 있는지, 그리고 자료 활용 시 유의할 점은 무엇인지에 대하여 개괄적으로 알아보았다. 자료를 공유하는 주요 관행들을 밝힘으로써 연구자가 필요로 하는 자료를 손쉽게 검색하여 취득하는데 도움을 주고자 했다. 지금 이 시간에도 새로운 데이터가 새로운 방법을 통해 공유되고 있으므로 모든 관행들을 적시한다는 것은 불가능한 일이다. 실제로 연구의 성격이나 자료공유에 대한 내부기준이 각 연구마다 다르기 때문에 원하는 데이터에만 접근 할 수 있게 하는 만능 키는 존재하지 않는다. 따라서 본 연구에서 제시한 방법들을 바탕으로, 오랜 시간 동안 많은 시행착오를 거쳐 경험을 축적하게 되면 원하는 데이터에 접근하는 자신만의 방식을 터득할 수 있을 것으로 생각된다. 2차 자료 사용이 처음인 연구자는 본 연구에서 제시한 대로 ICPSR과 미국 질병관리본부 홈페이지를 검색해 보면 주요 건강관련 데이터를 취득할 수 있을 것이다. 하지만 본 연구에서 제시한 방법은 궁극적인 해답이 아니라, 연구자의 연구목적에 더 적합한

데이터에 접근하기 위한 시작에 불과할 지도 모른다.

데이터 접근이 가능하고 그에 따른 적합한 유료 및 무료 통계분석 프로그램을 갖추었다면, 취득한 데이터를 어떻게 활용할 것인가 하는 문제가 남게 된다. 반복되는 내용이지만 2차 자료는 이론가설을 만들고(hypothesis generating study), 만들어진 가설을 검정(hypothesis test study)하는 데 유용하게 이용된다. 가설검정연구는 단계적으로 진행되는 것이 일반적이다. 즉, 특정 지역 및 특정 대상자에 국한된 이론을 정립(locally supported theory)한 다음에, 관련된 2차 자료를 통해 다른 지역 및 다른 인종에게도 이론을 확대적용시킬 수 있는지를 검토(globally supported theory)해야 한다. 예를 들면, 문인옥과 손은성(2004)은 우리나라 중소도시에 거주하는 고등학생들의 건강위험행위에 영향을 미치는 요인을 분석하였는데, 이는 특정지역에서만 성립될수 있는 이론일지도 모른다. 따라서, 미국 질병관리본부(CDC)에서 무료로 공유되고 있는 YRBSS(Youth Risk Behavior Surveillance System) 데이터를 이용하여 이론의 적용영역을 확대할 필요가 있다. 한국과 미국 청소년들의 건강위험 행위에 영향을 미치는 요인들에 차이가 있을 경우 그것이 문화적인 요인에 근거한 것인지, 혹은 사회 경제적인 요인 또는 유적적인 요인의 차이에 의해서 야기된 것인지를 밝힌다면 보다 보편적인 이론을 정립할 수 있을 것이다. 또 다른 예를 들자면, 이원재(2001)는 우리나라 몇몇 대학에 재학 중인 학생들을 대상으로 음주 양태를 분석하였다. 보다 더 보편타당한 이론정립을 위해서는 Harvard School of Public Health College Alcohol Study(Wechsler 등, 2000; Wechsler 등, 2002) 데이터를 국내 데이터와 함께 분석할 수 있다. 위의 예시에서 보는 바와 같이 이론영역의 확대를 위해서는 적합한 2차 데이터에 대한 접근이 무엇보다도 중요하다.

Health Level Seven (HL7)은 전자건강정보(electric health information)를 교환, 통합, 공유하기 위한 국제적인 표준화작업을 위한 기구를 의미한다(Blobel 등, 2009). 이러한 국제적인 표준화 노력은 연구와 연구간 혹은 국가와 국가간의 건강을 비교 분석하는 비교연구를 활성화시키고 있다(Ko 등, 2008). 최근 들어서는 특정 질병문제를 해결하기 위해 국가간 비교연구를 위한 컨소시엄을 조성하는 경우는 흔한 추세이다(Camille Ragin and Emanuela Taioli, 2009; Southam-Gerow 등, 2009). 이런 추세에 발맞추어 각종 저명 학술지에서도 국가간 비교연구에 대한 관심이 높은 것으로 알려졌다. 따라서, 2차 자료를 취득하여 국가간 비교연구를 한다면 학술지에 채택될 가능성이 높을 수도 있다.

본 연구에서 밝혀낸 자료공유를 위한 관행들을 바탕으로 인터넷상에 존재하는 2차 자료에 대한 접근성을 높이고 연구자 개인의 활용계획을 통해 연구를 진행한다면 많은 시간과 비용을 들이지 않더라도 훌륭한 학문적 성과를 이룩할 수 있으리라 사료된다. 지금도 인터넷에는 연구자의 손길을 애타게 기다리고 있는 청소년 건강관련 자료들이 무수히 존재한다. 그리고 이를 분석하기 위한 무료 통계프로그램 역시 무수히 존재한다. 자료가 없어서 혹은 통계프로그램이 없어서 논문을 쓰지 못한다는 것은 이제는 핑계에 지나지 않는다. 2차 자료에 대한 접근성 및 활용도를 높이기 위해 시간과 노력을 투자하는 것은 가치 있는 일이고 비용효과적이라 할 수 있다.

참고문헌

1. 문인옥, 손은성. 고등학생의 건강위험행위에 영향을 미치는 요인 -구리시 일부 고등학생을 중심으로-. 대한보건연구 2004;30(2):37-45.
2. 이원재. 대학생의 음주양태. 보건과 사회과학 2001;10:79-95.
3. Blobel B, Chronaki C, Stegwee R, Grain H, Hammond WE, Jaffe C, Kalra D, Koehn M, Macary F, Sabutsch S, Sheetham E, Schulz S, Tan M; HL7 Affiliates Council. HL7's comprehensive standards set and its international collaboration for enabling semantically interoperable eHealth and pHealth solutions. Stud Health Technol Inform. 2009;150:982-6.
4. Chandramohan D, Shibuya K, Setel P, Cairncross S, Lopez AD, Murray CJL, Basia Z. Snow RW, Binka F. Should Data from Demographic Surveillance Systems Be Made More Widely Available to Researchers? PLoS Med. 2008;5(2):e57.
5. Ko CY, Parikh J, Zingmond D. Secondary analyses of large population-based data sets: issues of quality, standards, and understanding. Ann Surg Oncol. 2008 Feb;15(2):395-6.
6. Piwowar HA, Becich MJ, Bilofsky H, Crowley RS, and on behalf of the caBIG Data Sharing and Intellectual Capital Workspace. Towards a Data Sharing Culture: Recommendations for Leadership from Academic Health Centers. PLoS Med. 2008;5(9):e183.
7. Ragin C and Taioli E. The Second International African-Caribbean Cancer Consortium for the study of viral, genetic and environmental cancer risk factors. Infect Agent Cancer. 2009;4(Suppl 1):S1.
8. Savage CJ, Vickers AJ. Empirical study of data sharing by authors publishing in PLoS journals. PLoS One. 2009 18;4(9):e7078.
9. Southam-Gerow MA, Hourigan SE, Allin RB Jr. Adapting evidence-based mental health treatments in community settings: preliminary results from a partnership approach. Behav Modif. 2009;33(1):82-103.
10. Wechsler H, Lee JE, Kuo M, Lee H. A College Binge Drinking in the 1990s: A Continuing Problem. The Results of the 1999 Harvard School of Public Health College Alcohol Study. Journal of American College Health 2000;48(5):199-210
11. Wechsler H, Lee JE, Nelson TF, Kuo M. Underage College Students Drinking Behavior, Access to Alcohol, and the Influence of Deterrence Policies: Findings From the Harvard School of Public Health College Alcohol Study. Journal of American College Health 2002;50(5):223-236.